Sequence effects in time trade-off valuation of hypothetical health states

Pinto-Prades, Jose Luis; McHugh, Neil; Donaldson, Cam; Manoukian, Sarkis

Published in:
Health Economics

DOI:
10.1002/hec.3942

Publication date:
2019

Document Version
Author accepted manuscript

Title: Sequence effects in time trade-off valuation of hypothetical health states

Abstract

Choice-based stated preference methods, such as Time Trade-Offs (TTOs), are used to establish health state utilities informing healthcare allocation. However, little is known about the presence of (position-dependent and precedent-dependent) sequence effects in the valuation of health states, despite techniques requiring respondents to evaluate several health states in a sequence. This paper is the first to explicitly test for the presence of sequence effects in the health domain using a new explanation based on contrast effects and preference imprecision. The implication being that randomisation cannot avoid sequence effects.

Six TTO questions were designed using the EQ-5D-3L descriptive system. These were grouped into two blocks of three and within each block four sequences were used. In an online survey, 1,197 Spanish respondents answered one grouping of three TTO questions. Results indicate that sequence effects can affect preferences as utilities of health states are biased downwards if preceded by a better health state and biased upwards if preceded by a worse health state.

This study informs our understanding of how context effects interact with preference elicitation methods, which is essential for interpreting survey results used to inform policy.

Keywords

Sequence effects, health state valuations, TTO, imprecision, contrast effects

1. Introduction

Most surveys conducted to calculate health state utilities ask subjects to evaluate several health states. For example, in EuroQol (Dolan et al., 1995; Dolan et al., 1996) subjects are asked to evaluate several health states, in a *sequence*, using the Time Trade-off (TTO) technique. However, in the valuation of health states very little is known about the presence of sequence effects. In principle, we would expect that calculating utilities of health state A and then of health state B would be the same (except for random error) as starting with B and then asking for A. Within stated preference tasks there is increasing evidence of the pervasiveness of sequence effects (Augestad et al., 2012; Day & Pinto Prades, 2010; DeShazo, 2002; McNair et al., 2011); yet in the health domain this issue is underexplored. One of the few papers to analyse this issue is Augestad *et al.* (2012) who find that mild health states have higher utilities if evaluated later in a sequence while the utilities of severe health states decrease. However, as Augestad *et al.* (2012) was not designed to study sequence effects it is difficult to know the reasons for their occurrence; although based on their results ex-post explanations can be offered. The aim of this paper is to explicitly test for the presence of sequence effects using a design based on a novel explanation of those effects.

Underpinning our explanation of sequence effects is the role of contrast effects. This is a psychological phenomenon that has been observed in multiple situations, beginning in psychophysics (Fernberger, 1920). More recently, it has been observed that contrast effects can affect judgements about social issues, for example, about self and others (Biernat et al., 1997), evaluation of physical attractiveness (Kenrick & Gutierres, 1980), food quality (Lahne & Zellner, 2015), happiness (Damisch et al., 2006) and economic

decisions (Simonsohn, 2006; Simonsohn & Loewenstein, 2006). While Fernberger (1920) noted a century ago that contrast effects could produce sequence effects[1], we add another element, namely, preference imprecision. We suggest that the intrinsic difficulty that people have in evaluating health states using methods, such as the Time Trade-Off (TTO) or the Standard Gamble (SG) can also contribute to the presence of sequence effects. Our proposal is that sequence effects can be explained by a combination of these two effects – contrast effects and preference imprecision. This theory is presented in Section 2. Based on this explanation/theory of sequence effects we designed a survey (Section 3) aimed at testing the predictions of the theory. We are not aware of any other study in the health literature explicitly designed to test for the existence of sequence effects. While "the typical way of neutralizing question order bias in the aggregate is to randomise the order in which different health states are valued" (Ternent & Tsuchiya, 2013, p545), if our explanation of sequence effects is correct, randomisation will not avoid those effects. Our results (Section 4) support this, as we present strong evidence of sequence effects which are not avoided through randomisation. Finally, we suggest in the Discussion (Section 5) that our results are not only relevant for the specific issue of sequences but have wider implications for the way that preferences are modelled in health economics.

In summary, our paper makes two main contributions:

- Theoretical contribution: the paper presents a new explanation of sequence effects in the health domain based on contrast effects and preference imprecision.

[1] "if a comparison stimulus, (……) immediately follows the lightest pair, for which, of course, the judgment is usually 'lighter,' there is a strong tendency that it will be judged 'heavier.' "(Fernberger, 1920, p.149).

- Empirical contribution: the survey is designed to test new hypotheses about sequence effects.

2. Theory

2.1 Defining sequence effects

Sequence effects can broadly occur in two different ways – position-dependent order effects and precedent-dependent order effects (Day et al., 2012). The former corresponds to the position of the objects in the sequence and the latter to the nature of the options in preceding tasks. For example, in the sequences 'A-B-C' and 'C-B-A' the position of B does not change but the preceding object evaluated changes. If object B is evaluated differently in both sequences this can be attributed to precedent-dependent order effects but not to position-dependent order effects. We next present a simple decision model that can help us to understand sequence effects.

2.2 Contrast effects and preference imprecision

Our explanation of sequence effects is based on the idea that responding to preference elicitation questions for health states is a complex task. In general, it is difficult for people to respond to preference elicitation questions when they are unfamiliar with the good, the task or both (Hausman, 2012; McFadden & Train, 2017); people are uncertain about their preferences. The evaluation of health states seems to be one of those situations. For example, when researchers conduct test-retest exercises of tasks, such as TTO, SG or Choice Experiments they usually find some element of variability in the responses (Feeny et al., 2004; Gamper et al., 2018; van Agt et al., 1994). In a recent

study, Gamper et al (2018) found that only 25% of subjects were perfectly consistent (i.e. repeated exactly the same choice between two options) in a choice experiment.

The complexity of the task has two implications for subjects' responses to TTO questions. One is that they can be subject to context effects. Simonson and Tversky (1992, p292) argue that "when people are uncertain about the values of options, they are more likely to use the context in determining the 'best buy'" and Tversky and Simonson (1993, p1184) state that context effects "are expected to vanish in situations where people have well-articulated preferences, and they are expected to be positive when the choice is more difficult and less certain". When it is difficult for people to know the precise value of an object, they tend to use relative comparisons to evaluate them. This can produce contrast effects. We can define a contrast effect as a negative (positive) change in the perception of an object prompted by recent exposure to a more positive (more negative) object. If the objects are health states and a subject considers that health state A is better than B, contrast implies that the perception of A will be better if B is evaluated first followed by A than if A is evaluated in the first place. The second implication is that there will be an element of variability in the evaluation of health states. Some people may change their responses to the same question from one moment to the next (Feeny et al., 2004; Gamper et al., 2018; van Agt et al., 1994). This leads to the next decision model.

2.3 Preferences, imprecision and contrast

Following Tversky and Simonson (1993) we assume that preferences can be represented as shown in equation 1:

1) $U_B(x) = u(x) + \beta f_B(x) + \Omega_x$

$U_B(x)$ is a context-dependent utility function where x denotes the object to be valued and the subscript indicates that the utility depends on the "background" context, namely, the effect of previous choices on the valuation of x. The utility of the object is a linear function of three elements, namely, u(x) is the context-free value of x, $f_B(x)$ is the effect of the background (i.e. contrast effects) and Ω_X is a random element. In our model Ω_X reflects within-subject variability. We can assume for the sake of simplicity, that Ω_X is a normally distributed random variable with zero mean.

We also hypothesize another characteristic of individuals' preferences will influence the evaluation of objects in a sequence, namely, people will try to be internally consistent when they evaluate several health states in a sequence. For example, they will try to avoid violating transparent dominance. This assumption is similar to the Coherent Arbitrariness effect observed by Ariely et al (2003). This characteristic of preferences will contribute to the generation of sequence effects, as we explain next.

2.4 Explaining sequence effects

We propose that preferences for health state S can be represented by a set (L^S) of potential utilities ($U_1^S, U_2^S, \ldots, U_n^S$) and it is *as if* the subject responds to a TTO question by choosing one of those utilities. If S is evaluated in the first position of the sequence, the response to the TTO will be influenced by the context-free utility of S and the random component Ω_X. If S is not the first health state in a sequence, the response to the TTO question will also be influenced by contrast, that is to say, by the severity of health states evaluated previously.

We can also assume that if two health states (A and B) are similar in terms of severity, the intrinsic utility will not be very different, contrast effects will not be very large and β will be small. However, there will be a lot of overlap between L^A and L^B produced by Ω_A and Ω_B. The opposite will happen if A and B are very different. If one is very mild and the other is very severe, contrast effects can be very large but there will be almost no overlap between L^A and L^B produced by Ω_A and Ω_B.

2.4.1 A sequence of two health states

Our model makes clear predictions regarding the direction of sequence effects. We start with the simplest case, namely, a sequence of two health states, X and Y. Assuming that X is perceived as better than Y (X≻Y) by the subject, U(Y) will be lower in the sequence X-Y than U(Y) in the sequence Y-X. The opposite will happen for X. In general, when a health state is evaluated in the second position of the sequence and it is preceded by a better (worse) health state, the utility will be lower (higher) than if it is evaluated first in the sequence. Those predictions are an immediate consequence of contrast effects, since Y will be perceived as more severe in the sequence X-Y than in sequence Y-X if X≻Y.

We also want to point out that uncertainty (the Ω_X) may also play a role, even if $E(\Omega_X)=0$, especially when the context-free U(X) and U(Y) are similar. In that case, there will be a lot of overlap between L^X and L^Y produced by Ω_A and Ω_B. In the sequence X-Y, the potential values of L^Y that the subject can use in her response, will be constrained by the value U(X) chosen from L^X since the subject wants to avoid violating dominance. For example, assume L^X is (0.50, 0.51…0.6) and L^Y is (0.45, 0.46…0.55). Assume that in the sequence X-Y, the subject responds U(X)=0.52. Since we assume that the subject wants to respect dominance, L^Y will be constrained to (0.45, 0.46…0.51) which will lead to U(Y)

being lower in the sequence X-Y than in the sequence Y-X. In the case of X, the prediction is the opposite. While both effects (uncertainty and contrast) work in the same direction to produce sequence effects, it seems logical to assume that contrast effects will be stronger when the two health states are very different since there will be a lot of contrast between the two. The role of uncertainty and internal consistency will be larger when the context-free utility of the health states are very close since, in that case, there will be a lot of overlap between L^X and L^Y. In summary, sequence effects could be the consequence of two effects. One, produced by contrast results in changes in perceptions and hence preferences, reflected through different valuations. The other, produced by imprecision affects valuations but not preferences due to a desire to be consistent in a limited valuation space.

2.4.2 A sequence of three health states

Assume now that we have three health states (the case we use in our study) that can be ranked by the subject from best to worst. Let us call 'B' the Best health state, 'W' the Worst and 'I' the Middle one. Predictions with three health states are more complicated since it is not clear how much "memory" a subject has. By "memory" we mean if the subject remembers or not, the response to the first question when responding to the third question or if it is only the previous response that influences her response to the third question. However, even if we do not know how much "memory" people have we can make some predictions. To explain that further, we will introduce what we call "Ascending", "Descending" or "Mixed" sequences. An Ascending sequence is a sequence where each health state is better than the previous one. In the case of our three health states, it would be W-I-B. The Descending sequence would be B-I-W and the rest would

be considered Mixed[2] sequences. Based on our previous model, predictions for Ascending or Descending sequences are clear but for Mixed sequences this is less so.

In the case of Ascending or Descending sequences, the effect goes in the same direction as if there were only two health states, but it is stronger. Assume we have Ascending sequence W-I-B. Our model predicts that U(B) in sequence W-I-B will be higher than in sequence I-B. The clearest reason is based on the consistency argument. This argument implies that L^I will already be constrained by the response to W, pushing U(I) upwards. This will further constrain the set of responses for B in L^B that the subject can use in order to maintain consistency. In relation to contrast effects, this will lead to a higher U(B) in sequence W-I-B if people compare B to the previous health state (I) after considering the initial health state (W). In the case of Descending sequences, B-I-W, the same arguments will lead to U(W) being lower than in the sequence I-W.

In summary, we have the following hypotheses when three health states are evaluated in a sequence (see columns 7 and 8 in Table 1):

H1: The best health state will receive higher values when evaluated in second or third position than when evaluated in the first position.

H2: The worst health state will receive lower values when evaluated in second or third position than when evaluated in the first position.

H3: The intermediate health state will receive lower values when evaluated after the best health state than when evaluated in the first position of the sequence.

[2] Mixed sequences are further subdivided into Mixed_1 and Mixed_2. These categories relate to the position of the best health state (second position in Mixed_1 and third position in Mixed_2).

H4: The intermediate health state will receive higher values when evaluated after the worst health state than when evaluated in the first position of the sequence.

H5: The utility of the best health state will be higher when evaluated in the third position of the sequence than when in the second position of the sequence, when the previous health state is the same in both cases.

H6: The utility of the worst health state will be lower when evaluated in the third position of the sequence than when in the second position of the sequence, when the previous health state is the same in both cases.

In the case of Mixed sequences, it is less clear the kind of prediction we can make for the health state evaluated in the third position. For this reason, we will abstain from making predictions for those cases.

3. Methods

3.1 Survey Design

TTO questions were designed using the EQ-5D-3L descriptive system[3] (Dolan et al., 1996). This instrument consists of five domains – mobility, self-care, usual activities, pain/discomfort and anxiety/depression – which have three possible levels – no problems (level 1), some problems (level 2) and extreme problems (level 3). Thus a health state 11111 refers to full health and 33333 refers to the worst health. This means the EQ-5D-3L defines 243 theoretically possible health states. Six different health states

[3] A five level EQ-5D descriptive system has now been developed (EQ-5D-5L) (Herdman et al., 2011).

were used from this instrument (see Table 1).

Health states were grouped in two blocks of three (Table 1). Block Φ: {W=22222, I=22211, B=11211} and Block Γ: {W=22322, I=22311, B=11311}. The blocks have clear relations of dominance (11211>22211>22222 and 11311>22311>22322). Four sequences within each of the two blocks generated eight groups, as seen in Table 1. These four sequences correspond to the sequences examined in our theoretical discussion: B-I-W (Descending), I-B-W (Mixed_1), I-W-B (Mixed_2) and W-I-B (Ascending). Each subject was allocated to one group and faced three TTO questions. This resulted in a within and between sample design.

Table 1 Survey Design and Hypotheses

Block	**Groups**	**Sequence**	**Sequence of TTO questions**			**Hypothesis: U health state in 2nd position vs. U same health state in 1st position**	**Hypothesis: U health state in 3rd question vs. U same health state in 1st position**
			1st	**2nd**	**3rd**	H: $U(B,I,W)_2$ vs. $U(B,I,W)_1$	H: $U(B,I,W)_3$ vs. $U(B,I,W)_1$
Φ {W=22222, I=22211, B=11211}	1	Descending	11211	22211	22222	H3: $U(I)_2 < U(I)_1$	H2: $U(W)_3 < U(W)_1$
	2	Mixed_1	22211	11211	22222	H1: $U(B)_2 > U(B)_1$	H2: $U(W)_3 < U(W)_1$
	3	Mixed_2	22211	22222	11211	H2: $U(W)_2 < U(W)_1$	H1: $U(B)_3 > U(B)_1$
	4	Ascending	22222	22211	11211	H4: $U(I)_2 > U(I)_1$	H1: $U(B)_3 > U(B)_1$
Γ {W=22322, I=22311, B=11311}	5	Descending	11311	22311	22322	H3: $U(I)_2 < U(I)_1$	H2: $U(W)_3 < U(W)_1$
	6	Mixed_1	22311	11311	22322	H1: $U(B)_2 > U(B)_1$	H2: $U(W)_3 < U(W)_1$
	7	Mixed_2	22311	22322	11311	H2: $U(W)_2 < U(W)_1$	H1: $U(B)_3 > U(B)_1$
	8	Ascending	22322	22311	11311	H4: $U(I)_2 > U(I)_1$	H1: $U(B)_3 > U(B)_1$

A choice-based procedure was used to estimate utilities for the different health states (an example of a choice is shown in Supplementary Materials 1). The first choice was

between 20 years in bad health (e.g. one of the three health states in a group) and 2 years in full health. This initial question was presented first as we wanted to know, as soon as possible, if the subject considered the health state as better or worse than dead. If the subject preferred 2 years in full health to 20 years in bad health, the second question was between 20 years in bad health and dead. In this way, after two questions we knew if the subject considered the health state as better or worse than dead.

If the subject preferred 2 years in full health to 20 years in bad health, she had to make four choices in a random order chosen by the computer, namely she had to choose between (20 years, bad health) and (6/10/14/18 years, Full Health). Since they had to respond to all four questions, chances are that subjects may have produced some inconsistency. In that case, subjects were shown a screen with all their responses and asked to resolve the inconsistencies. This choice process produced an interval of 2 (0-2 or 18-20) or 4 (2-6, 6-10, 10-14, 14-18) years where indifference should be located. This was further refined, via three (at most) additional choices, to an interval with a one-year range. A final open question asked subjects to state the number of months, within that one-year interval, at which they were indifferent.

If the health state was worse than dead, the indifference point was reached through a similar approach. It involved choices between immediate dead and the following profiles: (2 years, health state X; 18 years, Full Health; Dead), (6 years, health state X; 14 years, Full Health; Dead), (10 years, health state X; 10 years, Full Health; Dead), (14 years, health state X; 6 years, Full Health; Dead), (18 years, health state X; 2 years, Full Health; Dead). Again, the computer randomly presented five choices and subjects were invited to reconcile their inconsistencies. The preference interval was further narrowed

with (at most) three additional choices.

3.2 Data Collection

A market research company (Nexo S.L, Sevilla, Spain) was hired in June 2012. Initially the survey was delivered using face-to-face interviews but because of interviewer effects an online version was developed. This was piloted (n=200) before being delivered to the main sample. A sample of the Spanish population between 18 and 65 years of age (subjects over 70 years old were excluded as TTO questions included duration of 20 years which would exceed the average life expectancy of this age group) was recruited. Subjects were contacted via email and referred to the survey website. Incentives, in the form of points that are converted to goods, were used to encourage individuals to complete questionnaires.

The introduction to the survey outlined the study objectives and that it formed part of a research project for a Spanish university, it explained that we were interested in their perceptions of health problems and that there were no right or wrong answers. An example question came next involving a choice between 20 years in health state 22111 or 15 years in full health (11111); dead followed each choice. Each subject was randomised into a group and asked three TTO questions. The survey finished with a series of general socio-demographic questions.

3.3 Hypotheses

The majority of our sequence effects predictions are presented in Table 1 (columns 7 and 8). For example, in Group 1 health state 22211 (Intermediate) is in the second

position and this same health state is in position 1 in Group 2. Based on our theory, the utility of 22211 will be lower when evaluated in the second position than in the first position ($U_2<U_1$). Column 8 shows similar predictions for the health state evaluated in the third position relative to that evaluated first. The table shows how those predictions derive from our hypotheses.

While Table 1 can be used to understand H1-H4, hypotheses 5 and 6 are explained here:

- H5: $U(Best)_{Ascending\_sequence}>U(Best)_{Mixed1}$. This implies that U(11211) in Group 4 will be higher than U(11211) in Group 2. It also implies that U(11311) in Group 8 will be higher than U(11311) in Group 6.
- H6: $U(Worst)_{Descending\_sequence}<U(Worst)_{Mixed2}$. This implies that U(22222) in Group 1 will be lower than U(22222) in Group 3. It also implies that U(22322) in Group 5 will be lower than U(22322) in Group 7.

3.4 Data Analysis

3.4.1 Adjusting TTO scores

In TTO questions, health states 'full health' and 'dead' are assigned scores of 1 and 0, respectively. For health states valued better than dead, the value assigned to the health state (i.e. B, I or W) is x/20, where x equates to the number of years spent in full health. The value of health states considered worse than dead is calculated by –x/(20-x). Because the observed health state values ranged between 0 and -239, utilities <0 (approximately 10% of the sample) were normalised to -1 following Shaw et al. (2005), i.e. transformed values were obtained by dividing them by the lowest negative potential utility. Analysis is based on this normalised data.

3.4.2 Testing hypotheses

A linear regression model is used to test our hypotheses. The eight groups that respondents belong to are categorised according to our four different health state sequences: Ascending, Mixed_1, Mixed_2 and Descending (see Table 1). If the sequence does not matter, then the responses (utility values) should not depend on the type of group. Therefore, in a regression with the dependent variable being the utility of a health state (Best, Intermediate and Worst) the type of group should not systematically predict the utility level. The linear regression model for individual *i* is expressed in equation 2:

$$2) \quad U_i(\text{Health-State}_j) = \beta_1 \text{Grouptype}_x + \beta_2 \text{Grouptype}_y + \beta_3 \text{Grouptype}_z + \gamma_i X_i + \varepsilon_i \quad i=1\ldots N$$

Where β_1, β_2 and β_3 are the coefficients of interest, X_i is a vector of personal characteristics (Gender, Age, Marital status etc.), ε_i is a stochastic error term, j=1,2,3 (Health-State_1 = Best; Health-State_2 = Intermediate; Health-State_3 = Worst) and $\text{Grouptype}_{x,y,z}$ = Descending; Mixed_1; Mixed_2; and/or Ascending.

In the above equation, utility of a health state (e.g. U(Best)) as the dependent variable is regressed on group types to test the six hypotheses presented in Section 3.3. Using the utility of a specific health state as the dependent variable implies that each model includes one valuation for each individual. In each regression model, a group type is left out to satisfy the assumption of no multicollinearity. This excluded group type serves as a reference category and the β coefficients are always interpreted in relation to the

reference category. If there are no sequence effects the β coefficients should not be statistically different than zero; otherwise there are sequence effects and rejection or not of our six hypotheses will depend on the sign of the coefficient.

Robustness checks in the form of regressions on subsamples of subjects are also performed. First, we excluded the 20% fastest subjects to see how this impacted results as subjects who answered very quickly may not have taken the time to understand the TTO questions properly. Second, we excluded subjects who violated dominance (e.g. providing a higher utility to 22222 than 22211) at least once as this indicates subjects have been inconsistent with their responses (see Section 4.2).

4. Results

4.1 Sample

6,003 members of a market research panel were initially invited, by email, to participate in the survey in May-June 2013. 2,016 individuals (a 33.6% response rate) consented of whom 251 were randomly excluded as excess to quota[4]; 270 individuals did not complete the survey leaving a sample of 1,495 subjects to be randomly allocated to 10 groups. This study focuses on 8 of those 10 groups (n=1,197). Individual characteristics of the sample are shown in Table 2.

[4] Quotas were established according to sex and age (18-34, 35-55, 56-70).

Table 2 Individual Characteristics

Total (n)	1,197
Gender	
Female	50.38%
Male	49.62%
Age	
18-34	34.84%
35-54	45.28%
55-70	19.88%
Marital Status	
Married	58.06%
Single	34.59%
Other	7.36%
Education	
Primary school level or less	8.52%
Secondary school level	39.85%
Graduate level	51.63%
Employment Status	
Employed	59.06%
Unemployed	16.46%
Student	10.53%
Other	13.95%
Monthly Income	
Below 900 Euros	31.75%
901-1,500 Euros	30.16%
1,501-2,000 Euros	20.05%
2,001-3,000 Euros	11.53%
Over 3,000 Euros	6.52%
Survey Indicators	
Violation of dominance*	29.24%
Mean (standard deviation) survey completion time in minutes	15.59' (6.60')
*Refers to respondents who violated dominance (e.g. providing a higher utility to 22222 than 22211) at least once.	

4.2 Health State Utilities

Descriptive statistics relating to normalised health state utilities are shown in Table 3. A logical order, predicted by dominance, is shown for both means and medians. Medians remain the same when non-normalized utilities are used.

Table 3 Normalised health state utilities (medians and means)

Groups (Block)	Sequence	Order of questions			N	Means (Standard Deviations)			Medians		
		1st	2nd	3rd		U_1	U_2	U_3	U_1	U_2	U_3
1 (Φ)	Descending	11211	22211	22222	153	.813 (.240)	.649 (.342)	.437 (.423)	.896	.775	.496
2 (Φ)	Mixed_1	22211	11211	22222	151	.696 (.334)	.846 (.263)	.517 (.376)	.825	.904	.654
3 (Φ)	Mixed_2	22211	22222	11211	146	.722 (.311)	.583 (.368)	.907 (.212)	.796	.696	.946
4 (Φ)	Ascending	22222	22211	11211	157	.591 (.394)	.793 (.282)	.916 (.182)	.737	.896	.975
5 (Γ)	Descending	11311	22311	22322	147	.591 (.385)	.445 (.418)	.264 (.419)	.633	.496	.125
6 (Γ)	Mixed_1	22311	11311	22322	148	.509 (.456)	.717 (.372)	.350 (.484)	.694	.794	.492
7 (Γ)	Mixed_2	22311	22322	11311	148	.563 (.409)	.410 (.434)	.768 (.400)	.675	.496	.896
8 (Γ)	Ascending	22322	22311	11311	147	.442 (.416)	.619 (.346)	.734 (.329)	.542	.725	.846

4.3 Statistical tests of Sequence Effects

Table 4 shows the results of the regression analysis; only the coefficients of interest are presented (full regression results are shown in Supplementary Materials 2).

Table 4 Regression coefficients and associated hypotheses

Hypothesis	Dependent	Covariate[a]	**Coefficient**	**Standard**	**Pr > |*t*|**	**N**

	Variable			**Error**[b]		
H1	U(Best)[c]	Ascending	.136***	(.031)	.000	1,197
H1	U(Best)[c]	Mixed_1	.084***	(.031)	.006	
H1	U(Best)[c]	Mixed_2	.131***	(.031)	.000	
H2	U(Worst)[d]	Descending	-.181***	(.034)	.000	
H2	U(Worst)[d]	Mixed_1	-.070**	(.035)	.044	
H2	U(Worst)[d]	Mixed_2	-.037	(.034)	.271	
H3	U(Intermediate)[e]	Descending	-.055*	(.029)	.059	
H4	U(Intermediate)[e]	Ascending	.087***	(.027)	.002	
H5	U(Best)[f]	Mixed_1	-.053*	(.030)	.078	
H6	U(Worst)[g]	Mixed_2	.144***	(.034)	.000	

[a]All regression models controlled for gender, age, marital status, education level, labour market status and income level. See Supplementary Materials 2 for the full set of results. [b] Huber–White robust standard errors in parentheses. [c] Reference category: Descending. [d] Reference category: Ascending. [e] Reference category: Mixed. [f] Reference category: Ascending. [g] Reference category: Descending. *** Denotes significance at the 99% level, ** Denotes significance at the 95% level, * Denotes significance at the 90% level.

Overall, our results suggest the presence of sequence effects. H1-H2 show that the utility of the best (worst) health state (B) is higher (lower) when it appears in an ascending (descending) sequence than in a mixed sequence and these results are statistically significant. For example, in H1 U(Best) in an Ascending sequence is on average higher (0.136) than U(Best) in a Descending sequence (the reference category) and the result is statistically significant. The utility of the intermediate health state is also statistically significantly lower (higher) when it follows the best (worst) health state than when its evaluated first in the sequence (H3-H4). Finally, statistically significant evidence is provided for H5-H6 as the utility of the best (worst) health state is higher (lower) when evaluated third as opposed to second in the sequence, when the previous health state in the sequence is the same in both cases. Our robustness checks are consistent with these results (see Supplementary Materials 3 for robustness check results).

One interesting result that we did not predict but that helps the understanding of our data relates to the comparison of the sequences Descending vs. Mixed-1 and Ascending vs. Mixed-2 (see Table 3). In these comparisons the same health state is evaluated in third place (W in Descending and Mixed_1 and B in Ascending and Mixed_2) and they are preceded by the same health states but in reversed order. According to contrast effects, the role of the previous health state is larger the more similar it is to the health state being evaluated meaning there should be greater contrast for the health state evaluated third in the sequence I-B-W (Mixed_1) than in the sequence B-I-W (Descending). This should produce lower utilities for W in Mixed_1 than in the Descending sequence. However, the observed effect is the opposite; statistical analysis shows that U(W) is lower in the Descending sequence than in Mixed_1. Similarly, we also fail to observe the effect predicted by contrast for U(B) in Mixed_2 and Ascending; no differences in U(B) are observed. This suggests other effects, apart from contrast, are affecting our results. We believe it is imprecision. Imprecision predicts that we should find a lower U(W) in Descending than in Mixed_1 and a higher U(B) in Ascending than in Mixed_2. This occurs because in the sequence I-B-W there is very little overlap between L^B and L^W meaning this effect should be small. While in the sequence B-I-W there is more overlap between L^I and L^W which pushes U(W) lower. The same argument (in the opposite direction) applies to U(B). Our results suggest that in the case of bad health states (22222 and 22322) this second effect is stronger than contrast and that for the best health states (11211 and 11311) both effects may cancel each other out.

4.4 Sequence effects in 'simpler' preferences

Sequence effects are also observed in 'simpler' preferences. 'Simpler' preferences relate to two type of values. First, whether health states are considered better or worse than

dead for which subjects should have more defined preferences; this test helps us explore whether eliciting preferences using an internet survey may be causing our results. There are two reasons for this. One is that the choice between (Health State X, 20 years) and dead, was the second choice they saw, after the choice between (Health State X, 20 years) and (Full Health, 2 years). Thus subjects should not be confused or tired by this stage. The second reason is that this question is very clear cut - you prefer immediate dead or not. The effect is very impressive especially for the worst health states in each block. In the case of 22222, the proportion of subjects who consider that the health state is worse than dead moves from 8.9% when it is evaluated in the first position to 19.0% in the Descending sequence. While in the case of health state 22322, the percentages are 19.0% when evaluated in the first position compared to 29% in the Descending sequence. The second 'simpler' preference relates to what we have called "extreme traders" - those who gave up one month (the minimum) out of 20 years in order to improve quality of life. This time we focus on the best health states, since this effect mainly affects the better health states. We find that 18.3% are "extreme traders" if 11211 is evaluated in the first position and this jumps to an impressive 44.6% in the Ascending sequence. In the case of health state 11311 the percentages are 8.2% when evaluated in the first position and 21.1% in the Ascending sequence.

5. Discussion

Choice-based stated preference methods, such as TTOs and SGs, are used to establish health state utilities that inform decisions of national Health Technology Assessment (HTA) agencies regarding the allocation of scarce healthcare resources. The assumption of procedural invariance – irrelevant changes to the order in which health states are

evaluated will not alter their value – underlies these methods. However, our results question the validity of this assumption. Evidence is provided that sequence effects, can affect preferences. Specifically, utilities of health states are biased downwards if preceded by a better health state and biased upwards if preceded by a worse health state. Additionally, our results suggest randomisation alone will not make these effects disappear; precedent-dependent order effects will still occur even if we randomize. We explain these results using a model that recognises that preferences in a TTO or SG questions are context-dependent and imprecise.

It is important to clarify that we do not think that all studies where utilities for health states are estimated in a sequence will show the strong effects observed in this study. Some of the features of our design may exaggerate these effects. For example, having only three health states makes it easier for people to compare between health states and to try and be consistent. However, the purpose of our study design was not to minimize sequence effects but to try and understand them. The results of our study mean we better understand the process by which those effects can happen; that, for example, can help us explain the results of Augestad et al (2012) since they obtain, in a less stylized design, the same results that our model predicts. Having clarified this point, we move to the implications of this study.

In addition to aiding the understanding of sequence effects our study adds to the evidence base of studies that have observed preference elicitation procedures that should (under basic rationality assumptions) be equivalent, produce very different results. The reaction of researchers to such results sometimes consists of suggesting other kinds of methods or techniques that could solve these problems. By “solve” we

mean methods that elicit "true" values. While we do not reject the need to develop better methods, the problem is that most of those models are based on a rational model where a subject's response reflects true preferences. Yes, an error term, is added to the utility function but since it is assumed to have a zero mean a large sample size is thought to be enough to find out true preferences. Our model suggests that even if this error has zero mean, it can generate biases. This paper suggests that the intrinsic difficulty in responding to TTO questions may produce context-dependent and imprecise preferences. This generates problems for all kind of preference elicitation methods that assume subjects reveal context-free values for health state values except for the influence of a random component (see Bansback et al. (2012) as an example) and that when aggregated do not produce any bias. However, if the model we have used to explain our data describes people's preferences better than the standard model, in the case of the evaluation of health states, even the best methods may not provide that "true" value. This is important because if preferences are imprecise, and if subjects make relative comparisons when they respond to preference elicitation questions, we need to understand those effects in order to separate out the true component of preferences from the influence of other elements in subjects' responses.

6. Conclusion

We observe sequence effects in the valuation of health states that are in line with predictions arising from a model that incorporates preference imprecision. It is suggested these effects will not disappear with randomisation. Understanding how

preference imprecision and relative comparisons interact with the methods used to elicit preferences is essential for interpreting the results we get from surveys.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*(1), 73-106.

Augestad, L. A., Rand-Hendriksen, K., Kristiansen, I. S., & Stavem, K. (2012). Learning effects in time trade-off based valuation of EQ-5D health states. *Value in Health, 15*(2), 340-345.

Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics, 31*(1), 306-318.

Biernat, M., Manis, M., & Kobrynowicz, D. (1997). Simultaneous Assimilation and Contrast Effects in Judgments of Self and Others. *Journal of Personality and Social Psychology, 73*(2), 254-269.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied, 12*(3), 166.

Day, B., Bateman, I. J., Carson, R. T., Dupont, D., Louviere, J. J., Morimoto, S., . . . Wang, P. (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management, 63*, 73-91.

Day, B., & Pinto Prades, J. L. (2010). Ordering anomalies in choice experiments. *Journal of Environmental Economics and Management, 59*, 271–285.

DeShazo, J. R. (2002). Designing Transactions without Framing Effects in Iterative Question Formats. *Journal of Environmental Economics and Management, 43*, 360-385.

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1995). *A social tariff for EuroQoL: Results from a UK general population survey*. Discussion Paper No. 138. Centre for Health Economics, University of York.

Dolan, P., Gudex, C., Kind, P., & Williams, A. (1996). The time trade-off method: results from a general population study. *Health Economics, 5*(2), 141-154.

Feeny, D., Blanchard, C. M., Mahon, J. L., Bourne, R., Rorabeck, C., Stitt, L., & Webster-Bogaert, S. (2004). The stability of utility scores: test-retest reliability and the interpretation of utility scores in elective total hip arthroplasty. *Quality of Life Research, 13*(1), 15-22.

Fernberger, S. W. (1920). Interdependence of judgments within the series for the method of constant stimuli. *Journal of Experimental Psychology: Applied, 3*(2), 126-150.

Gamper, E.-M., Holzner, B., King, M. T., Norman, R., Viney, R., Nerich, V., & Kemmler, G. (2018). Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States. *Value in Health, 21*(8), 958-966.

Hausman, J. (2012). Contingent Valuation: From Dubious to Hopeless. *Journal of Economic Perspectives, 26*, 43-56.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M. F., Kind, P., Parkin, D., . . . Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research, 20*(10), 1727-1736.

Kenrick, D. T., & Gutierres, S. E. (1980). Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology, 38*(1), 131.

Lahne, J., & Zellner, D. A. (2015). The great is the enemy of the good: Hedonic contrast in a coursed meal. *Food Quality and Preference, 45*, 70-74.

McFadden, D., & Train, K. (2017). *Contingent Valuation of Environmental Goods*: Edward Elgar.

McNair, B. J., Bennett, J., & Hensher, D. A. (2011). A comparison of responses to single and repeated discrete choice questions. *Resource and Energy Economics, 33*, 554-571.

Shaw, J. W., Johnson, J. A., & Coons, S. J. (2005). US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical Care*, 203-220.

Simonsohn, U. (2006). New Yorkers Commute More Everywhere: Contrast Effects in the Field. *The Review of Economics and Statistics, 88*(1), 1-9.

Simonsohn, U., & Loewenstein, G. (2006). Mistake #37: The effect of previously encountered prices on current housing demand. *The Economic Journal, 116*(508), 175-199.

Simonson, I., & Tversky, A. (1992). Choice in Context: Tradeoff Contrast and Extremeness Aversion. *Journal of Marketing Research, 29*(3), 281-295.

Ternent, L., & Tsuchiya, A. (2013). A Note on the Expected Biases in Conventional Iterative Health State Valuation Protocols. *Medicial Decision Making, 33*(4), 544-546.

Tversky, A., & Simonson, I. (1993). Context-Dependent Preferences. *Management Science, 39*(10), 1179-1189.

van Agt, H. M., Essink-Bot, M.-L., Krabbe, P. F. M., & Bonsel, G. J. (1994). Test-retest reliability of health state valuations collected with the EuroQol questionnaire. *Social Science and Medicine, 39*(11), 1537-1544.